DOCUMENT RESUME

ED 346 113 TM 018 331

AUTHOR Masters, James R.
TITLE A Study of Arbitrations in Pennsylvania's Writing Assessment.
PUB DATE Apr 92
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Arbitration; Decision Making; *Educational Assessment; Essay Tests; Evaluation Problems; Grade 6; Grade 9; High Schools; *Holistic Evaluation; Intermediate Grades; Interrater Reliability; Problem Solving; Prompting; *Scoring; *State Programs; Student Evaluation; Testing Programs; *Writing Evaluation
IDENTIFIERS Direct Assessment; *Pennsylvania

ABSTRACT

In 1991 Pennsylvania began implementation of a direct writing assessment at the sixth-grade and ninth-grade levels. A total of 18,758 sixth graders and 16,575 ninth graders wrote a response to 1 of 9 prompts reflecting 3 modes of writing. A six-point holistic scale was used to score the papers, with two readers scoring each paper. A third reader, or arbitrator, helped resolve any disagreements in the two scores. In all, there were 1,874 arbitrations. This study focused on the reasons for these arbitration decisions, asking whether such factors as the particular prompt, the writer's grade level, or the achievement level of the readers were related to such disagreements. Results indicate that arbitration rates varied by prompt, by mode, and for each of the two grade levels. Approximately 50% of the raters showed a tendency toward giving either high scores or low scores. At one-half of the tables of raters, one individual was found to be involved in at least 40% of the arbitrations. Neither characteristics of the raters nor the achievement level of the raters' students were found to be related to the arbitration statistics computed for each rater. There are 7 tables of study findings and an 11-item list of references. An appendix contains the holistic scoring guide. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JAMES R. MASTERS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346113

A STUDY OF ARBITRATIONS IN PENNSYLVANIA'S WRITING ASSESSMENT

James R. Masters

Pennsylvania Department of Education

Presented at the Annual Meeting of the
National Council on Measurement in Education

San Francisco, California
April 1992

TM018331

BEST COPY AVAILABLE

ABSTRACT

In 1991 the state of Pennsylvania began implementation of a direct writing asssessment at the sixth and ninth grade levels. Students were asked to write a response to one of nine different prompts reflecting three modes of writing. A six-point holistic scale was used to score the papers, with two readers scoring each paper. A third reader, or arbitrator, helped resolve any disagreements in the two scores. This study focused upon the reasons for these arbitration situations. It was asked whether such factors as the particular prompt, the writer's grade level or the achievement level of the readers' students were related to such disagreements. The results indicated that arbitration rates varied by prompt, by mode and for each of the two grade levels. Approximately 50 percent of the raters showed a tendency toward giving either high scores or low scores. At one-half of the tables of raters, one individual was found to be involved in at least 40 percent of the arbitrations. Neither characteristics of the raters nor the achievement level of the raters' students were found to be related to the arbitration statistics computed for each rater.

INTRODUCTION

In recent years many states have instituted a direct writing assessment in their school districts as a means of examining their students' abilities to write compositions. The state of Pennsylvania began such an assessment in the spring of 1991 when grade 6 and 9 students from 88 public school districts and 39 nonpublic schools participated.

A Writing Assessment Advisory Committee, composed of educators from school districts and universities from all parts of the state, helped make the major decisions necessary to this implementation. This committee had met for two years prior to the first assessment to mold it into a desirable one for the state. One of the products of their efforts was a "Writing Assessment Handbook" (Peters and Miller, 1990), produced to describe the program to district personnel.

Among the decisions made by the committee was the use of a six-point holistic scale as the vehicle for obtaining student scores. This scale (shown in the Appendix) was developed by the committee to reflect characteristics of effective writing, grouped under the five headings of Focus, Content, Organization, Style and Conventions. To obtain the highest score on this scale, a "6," a student's paper would have to show evidence of being at a high level on all five characteristics.

In the 1991 assessment, each paper was scored independently by two readers, using the holistic scale. In Pennsylvania's assessment process, scores are considered consistent which are exactly the same or different by only one point. Those which differ by more than one point are considered inconsistent and this inconsistency must be resolved by a third reader called an arbitrator.

Papers which go to arbitration can be viewed as instances of differences of opinion between readers about the quality of writing in a particular paper. However, the training of readers is an attempt to objectify the rating process so that a common scoring scale is internalized. Papers which must be arbitrated are evidences of the inability of raters to utilize the scoring scale in exactly the same way and are thus indications of unreliability in scoring.

This study was undertaken to investigate the arbitrations which took place in the first year of Pennsylvania's writing assessment. The intent was to learn about these as a means of reducing their incidence in future assessments.

REVIEW OF RELATED RESEARCH

Large-scale direct writing assessment as it is now practiced is of relatively recent origin. Because of this, not a great deal of research of this process has been performed. However, over the years much research has addressed a variety of aspects of the composition process.

Two major types of investigations are most relevant to the present one. These are studies which have dealt with characteristics of raters which may influence their ability to serve as reliable scorers and studies which have investigated features of the rating situation which may influence the accuracy of the process.

One relevant variable which has been researched is the ability of raters to put aside their inherent biases and internalize the scoring rubric being used. Huot, in summarizing the research in this area concluded: "Reader expectation shaped by personal and professional experience will always be a strong yet hard to define influence on holistic raters. Assessment procedures work to control the natural variability present in the fluent reading of individuals who have had different experiences and therefore possess a range of expectations as readers". (1990, p. 255).

Hake (1986) recognized this problem as one which had to be dealt with and calibrated her raters on a scale from "lenient" to "harsh." She argued that requiring raters to work toward agreement in becoming trained as holistic scorers could cause valuable responses to students' texts to be lost. Stock and Robinson (1987) concluded that variability in the rating of student essays is inevitable because this is part of the normal process of reading. They stated that readers react differently to text materials, and raters are not different in this respect from other readers. Webb, et al (1990) faced the same type of problem in a different context, that of an oral examination involving ratings

of the performance of a large number of persons. They concluded that rater stringency/leniency, or bias, was a reality, but could be dealt with. If it was not possible to control it through balancing the situation so that all raters observed all those involved in the performance sessions, statistical procedures could be used to carry out controls. However, they believed rater inconsistency was a more difficult problem to deal with, one that probably was best handled through providing effective training prior to the rating task.

Research addressing important aspects of the rating situation has focused upon a variety of factors. One factor which has been found to be important is related to the expectation problem described above. In this case, however, the expectations inherent in raters are brought out and made to interfere with the process because of some aspect of the rating situation which has been set up. As one instance of this type of interference, Diederich (1974) found that raters scored the same essays higher when they were informed that these came from honors classes. Similarly, Rigsby (1987) reported that in departmental competency examinations higher scores on essays were obtained when it was made known that these papers were written by upperclassmen.

In another study which addressed aspects of the rating situation which could influence the accuracy of ratings, Freedman (1981) attempted to determine the impact upon raters of three types of factors: essay variables, reader variables and environmental variables. A total of 64 essays written on two types of topics (opinion and quotation) were used in the study. Scoring partners, trainers and topic type were rotated for raters. Freedman concluded that the essay variable contributed most to the scores of raters. The only other variable which seemed to influence rater perception of writing quality was the particular trainer involved.

Daly and Dickson-Markman (1982) studied the impact upon the rating of essays of the sequence of papers read. They questioned, for example, whether an essay read after a series of excellent ones would be rated lower than would have typically been the case. They set up a variety of sequences and studied their impacts upon a control paper. They did find that sequence affected ratings. As one instance of this, they found that a control paper read after a series of four low quality papers was rated significantly higher. Quellmalz (1984) argued, similarly, that the mix of

papers read by raters was important. She stated that it was an effective procedure to have different grade levels write to the same prompt and have all these papers scored together. She maintained that scoring each grade level separately could lead to a norming within grade level effect, with the same scale point meaning different things at different grade levels. Meredith and Williams (1984) wrote that statements made to raters during the rating process could have a major impact in their reaction to the scale. For example, telling a rater he or she is scoring too low could result in a radical swing of his or her internal scale.

Thus, some research has been accomplished which has addressed similar variables to those studied in the present investigation. However, no study was found which examined this area utilizing arbitrations in the same way as in the present study.

METHOD

Nine prompts were used in the 1991 Writing Assessment. Three of these were Narrative in mode, three were Informational and three were Persuasive. These prompts were chosen as the best of the 20 tried out in the spring of 1990 with over 28,000 sixth and ninth graders from 232 schools. All nine prompts were used with both sixth graders and ninth graders.

The 1991 assessment took place in voluntary districts and nonpublic schools. The purpose was not to test the abilities of individual students, but rather to give schools information about the strengths and weaknesses of their writing programs. A total of 18,758 sixth graders and 16,575 ninth graders wrote essays. They were given two 40-minute periods in which to write both a first draft and a final essay. Each student responded to only one of the nine prompts. Students were not given a choice of prompts, but rather responded to whichever one was randomly distributed to them.

All essays were sent to a central location for scoring. The majority of the readers who met at this location were from the schools participating in the assessment. The scoring of papers for each

mode took place in a separate room. Chief Readers conducted training in each of the rooms. For each mode six anchor papers were used in this training as examples of ones which would be expected to receive each of the six scores.

In all three rooms either seven or eight readers and a Table Leader sat at each table and constituted a scoring group. In each room there were eight tables of eight or nine people. The Table Leader distributed papers for first readings, for second readings and for arbitrations. In some cases the Chief Reader served as a second reader and often as an arbitrator.

In order that each paper would be read under the same conditions, those from sixth graders were mixed in with those from ninth graders and papers from all schools involved were similarly mixed. Raters were not provided any information about the grade level, sex, race or school of the writer of any paper.

The arbitration data for the study came from Table Leaders' records of all arbitrations which had been carried out. These included a listing for each arbitration of which two raters had disagreed, what scores they had given, who the arbitrator had been and what score this person had given. In addition, the paper number for which the score disagreeement occurred had been recorded. It was therefore possible to match the paper number on the arbitration summary with that on the individual student computer file and thus to learn additional information about the student, such as his or her school district and building.

From the data available it was possible to compute the number of arbitrations for each rater and to sum these for all raters at the same table to determine table totals. However, one limitation of the study was the lack of any information about the overall performance of each rater or each table. It was known how many papers were read by the entire room of raters, but it was not known how many papers were read by each rater or each table. On the positive side, it was known that each rater worked for the same amount of time and that there was an attempt to spread the work around so that it was equally distributed across tables and for individual raters within tables.

Because this spreading around was not a perfect process, any analysis of the study which involves individual or table level results is not precise.

The study which was performed was exploratory in nature. The overall intent was to learn as much as possible about the arbitration process. More specifically, a series of questions about the arbitrations was asked and answered through examining the data of the study. In general, it was asked what had led raters to disagree about the papers they had read. Three major sources were investigated--the prompts used, characteristics of the raters and characteristics of the rating situation which had impacted upon the process.

RESULTS

Analyses Involving Prompts, Modes and Scores

A total of 1874 arbitrations took place for the 35,333 papers which were scored. This is an overall arbitration rate of 5.3 percent. To perform the major analyses of the study there was a need to match the identification numbers of the arbitrated papers with those on the file of scored papers. Because of errors on the part of Table Leaders in recording some of these for the arbitrated papers, 52 could not be matched. These were deleted and the number to be analyzed in the study was therefore reduced to 1822.

Table 1 shows the arbitration results for each grade level. These are listed for each prompt and mode. Also shown is the mean score for each prompt and mode given by all raters who scored this mode. The scores which entered into these means were the sums of the scores of the two raters who read each paper. Descriptions of each prompt could not be included due to a need to maintain some degree of security over their subjects.

Table 1
Arbitration Rates and Mean Total Scores for
Each Prompt, Each Mode and Overall

	Grade 6		Grade 9	
Prompt Number	Percent of Papers Arbitrated	Mean Total Score	Percent of Papers Arbitrated	Mean Total Score
1	5.71	6.74	5.27	8.12
2	5.44	6.59	5.19	8.13
3	6.28	6.60	5.61	8.06
Total for Narrative Mode	5.81	6.64	5.36	8.10
4	2.78	5.88	5.58	7.32
5	4.56	5.82	6.09	7.50
6	3.11	5.74	5.06	7.14
Total for Informational Mode	3.48	5.81	5.58	7.32
7	5.44	5.95	5.15	7.66
8	6.07	5.91	5.93	7.47
9	4.49	5.96	5.36	7.40
Total for Persuasive Mode	5.34	5.94	5.48	7.51
Overall Total for All Prompts	4.88	6.14	5.47	7.65

As shown in Table 1, arbitration rates were found to differ by grade level, by mode and for individual prompts. At the grade 6 level there was, in general, a direct relationship between the arbitration rate and how well students scored.

The Informational mode tended to be both the hardest for grade 6 students to write to and the lowest in arbitration rate. The prompt with the lowest grade 6 arbitration rate was number 4, "Special Person," for which students were asked to describe a special person in their lives.

Grade 6 students found it least difficult to prepare a Narrative essay. The arbitration rate for this mode was more than 2 percent higher than that for the Informational mode, however. The prompt for which this rate was highest was number 3, "Time Machine," for which students were asked to describe what they would predict life to be like fifteen years in the future.

As would be expected, grade 9 students' essays were rated higher than were those of grade 6 students. Grade 9 students experienced their greatest difficulty in producing an Informational essay and their greatest success in writing to a Narrative prompt. However, there was little evidence of any relationship between scores achieved for a prompt and the arbitration rate for it. In fact, although there were differences in arbitration rates across prompts, there was very little difference across modes. The prompts for which the grade 9 arbitration rates were highest and lowest were both Informational. The prompt evidencing the lowest rate was number 6, "Holiday Explained," for which students were asked to provide a description of a holiday of their choice. The prompt with the highest rate was number 5, "My Hero/Heroine," for which students were directed to write an essay about someone they admired.

Table 2 provides a summary of the numbers and percentages of each score which were given at each grade level. Also shown is the percentage of each score which went to arbitration. It should be noted that in this table the total number of scores is doubled, since it reflects the scores of both raters who read a paper.

Table 2

Numbers of Each Score Given and the Percent of Each Which Went to Arbitration

	Grade 6			Grade 9		
Score	Number Given	Percent Given	Percent of Those Given Which Went to Arbitration	Number Given	Percent Given	Percent of Those Given Which Went to Arbitration
0	8	0.02	25.00	1	0.00	0.00
1	1696	4.56	9.43	414	1.26	9.42
2	8588	23.09	4.61	2740	8.36	10.00
3	14461	38.87	2.88	8826	26.92	4.84
4	9594	25.79	4.67	11792	35.97	3.64
5	2490	6.69	12.65	6969	21.26	5.83
6	363	0.98	25.07	2042	6.23	11.70

It can be seen first of all in the table that the most common score at the grade 6 level was a 3, given to approximately 39 percent of the papers read. The next most common scores at this level were 2's and 4's, both given to approximately one-fourth of the papers read. At this level only about 7 percent of papers were given a score of 5 and less than one percent were judged to be worthy of a 6.

At the grade 9 level, a 4 was the most common score, given to approximately 36 percent of the papers scored. Next most common was a 3 given to 27 percent of papers and a 5 given to 21 percent of papers. At this level, 6 percent of papers were judged to be of 6 quality and only 1 percent were felt to be poor enough to warrant a 1.

The arbitration data provide an interesting picture of the differences across grade levels. At the grade 6 level, fully 25 percent of the 6's which had been given went to arbitration, and an additional 12.7 percent of 5's were arbitrated. Thus, close to 40 percent of the high scores given by one rater of a grade 6 paper were disputed by another rater. In addition, over nine percent of the very low ratings, the ratings of 1, went to arbitration.

Readers of grade 9 papers differed most often when one person gave either a very high rating or a very low rating. Approximately 19 percent of 1's and 2's were arbitrated and approximately 16.5 percent of 5's and 6's were arbitrated.

Table 3 adds to the picture by showing the numbers and percentages of pairs of scores which were arbitrated. At the grade 6 level, the most common disagreement (38 percent) was a 2 versus a 4. In these cases one reader felt the writer had produced a competent paper and the other person found relatively major deficiencies with it. The second most common disagreement at this level was a 3 versus a 5 (29 percent). In these cases both readers agreed that the paper was near the competent level, but one reader believed it to be below competence and the other found it to be above this.

Table 3
Percentages of Each Pair of Scores
Which Went to Arbitration

Two Scores Given	Grade 6		Grade 9	
	Frequency	Percent	Frequency	Percent
0-2	2	0.22	0	0.00
1-3	129	14.13	30	3.33
1-4	29	3.18	7	0.78
1-5	2	0.22	2	0.22
2-4	348	38.12	223	24.75
2-5	46	5.04	46	5.11
2-6	0	0.00	4	0.44
3-5	266	29.13	358	39.73
3-6	21	2.30	36	4.00
4-6	70	7.67	199	22.09

The two most common disagreements at the grade 9 level were the same as those which occurred for grade 6, but their order of occurrence was opposite to that for grade 6. Approximately 40 percent of disagreements about grade 9 papers were of the 3-5 variety and 25 percent were 2-4 disagreements. In addition, 22 percent of grade 9 disagreements were 4-6 in nature. In these cases one reader saw the paper as merely competent; the other found it to be exceptional.

Table 4 was prepared as a final way of looking at the arbitrations which had been performed. This table summarizes what happened when an arbitrator read the same paper as the two original readers.

One of the most interesting results shown in this table is at the grade 6 level for arbitrations in which one of the original scores was a 6. A total of 91 such cases occurred (i.e., 21 arbitrations of the 3-6 variety and 70 arbitrations of the 4-6 variety). In only 7.1 percent of cases did the arbitrator agree that a 6 was warranted. Thus, for only 6 of the 91 papers did the arbitrator also give a 6 rating.

Table 4
Arbitrator Scores Given for Each Pairing
of Scores Which Went to Arbitration

	Grade 6							Grade 9						
Two Scores Given	Number Which Occurred	% of Each Score Given by Arbitrator 1	2	3	4	5	6	Number Which Occurred	% of Each Score Given by Arbitrator 1	2	3	4	5	6
0-2	2	100.0						0						
1-3	129	8.5	53.5	35.7	2.3			30	16.7	33.3	40.0	10.0		
1-4	29	13.8	24.1	37.9	24.1			7	14.3		42.9	28.6	14.3	
1-5	2		50.0		50.0			2			50.0		50.0	
2-4	348	0.9	25.0	49.7	22.4	2.0		223	0.9	14.3	49.8	31.4	3.6	
2-5	46		21.7	37.0	39.1	2.2		46		8.7	34.8	37.0	19.6	
2-6	0							4				25.0	75.0	
3-5	266		2.3	33.1	50.0	13.2	1.5	358			23.7	51.7	22.6	2.0
3-6	21			9.5	47.6	42.9		36			5.6	36.1	41.7	16.7
4-6	70			8.6	42.9	41.4	7.1	199			6.0	33.7	45.7	14.6

In general, Table 4 provides a kind of validation of the process being used to score papers. For the situations which occurred most often, those for which the two raters differed by two points, the score given most often by the arbitrator was that which fell between the two original scores. In effect, the arbitrator tended to provide an average rating, to view the paper in a less extreme way than the two original scorers. Also, for the most common types of disagreements, there is evidence in Table 4 that the great majority of the time the arbitrator saw the paper as one which fell in quality somewhere along the score interval between and including the two scores.

Rater Tendencies Toward High or Low Scores

One set of analyses of the study focused upon whether any patterns of response for individual raters could be found in the arbitration data. Table 5 summarizes the results of an investigation of the extent to which particular raters tended to give either the higher or the lower score of the two which were arbitrated.

Table 5
Numbers/Percentages of Raters with Tendencies to Give Either the Higher or the Lower Score of the Pair of Scores Being Arbitrated *

Mode	No. of High Raters	No. of Low Raters	Raters with Either a High or a Low Tendency		Raters Showing No Tendency		Total Number of Raters
			No.	%	No.	%	
Narrative	15	17	32	50.79	31	49.21	63
Informational	18	18	36	57.14	27	42.86	63
Persuasive	17	11	28	44.44	35	55.56	63
All Modes	50	46	96	50.79	93	49.21	189

* Raters characterized as having a "high" rating tendency were those who, at least 2/3 of the time, had given the higher of the two scores being arbitrated. Raters characterized as having a "low' rating tendency had given the higher of the two arbitrated scores 1/3 or less of the time. All other raters were characterized as persons who "showed no tendency" toward either a high or a low pattern of ratings.

To prepare this table a determination was made for each rater of the percentage of time he or she gave the higher score of the two. Those who did this at least two-thirds of the time were arbitrarily defined as having a "high" rating tendency and those who did this one-third or less of the time were arbitrarily defined as having a "low" tendency. All those whose percentages fell between one-third and two-thirds were considered not to have a rating tendency.

As can be seen in Table 5, about 50 percent of the raters who participated in the assessment were found to have a rating tendency as defined above. The number who tended to respond with high ratings was very similar to the number who tended to respond with low ratings. High or low tendencies occurred most often among those scoring papers in the Informational mode and least often among those scoring papers in the Persuasive mode.

It should be made clear at this point that the tendencies being described were only with respect to the arbitrations. It would be expected that, in general, raters with high tendencies would have high overall mean ratings, but no data were available to confirm this.

In an attempt to understand as much as possible about the rating tendencies which were found, two further analyses were performed. These were descriptive in nature; each will be outlined briefly.

First, it was asked whether those who had exhibited some tendency had begun their rating task in this way or whether there had been some movement toward the tendency over time. To perform this analysis the tendency percentages described above were computed for each rater for the first one-third of the rater's arbitrations, for the second one-third and for the last one-third. The results indicated that the majority of those with tendencies had shown these at the outset of the scoring session. Of the 48 raters who had exhibited low tendencies, 32 (69.6 percent) had shown this tendency during the first one-third of their arbitrations. Of the 50 raters with high tendencies, 41 (82 percent) had begun with this same tendency. Overall, then, 73 out of 96 (76 percent) of those with rating tendencies had begun their ratings with the same tendency they had shown during the major portion of their assignment.

Finally, it was asked whether those who had shown some rating tendency had disagreed most with raters who had shown the opposite tendency. In effect, it was asked whether a major reason for having to incorporate an arbitration process into the rating system was that high raters and low raters disagreed, not because of the particular papers being scored, but because of their inherent tendencies to view all papers differently from each other.

This was not found to be the case. Of the 46 low raters, 18 (39.1 percent) disagreed most often with high raters. Of the 50 high raters, 17 (34 percent) disagreed most often with low raters. Thus, only 35 out of 96 (36.5 percent) of those who had shown some rating tendency disagreed most often with those who had shown the opposite rating trend.

The Numbers of Arbitrations for Individual Raters

It was obvious from examining the numbers of arbitrations which had taken place for each rater that a great deal of variance existed. All raters had been involved in at least two arbitrations. The highest number for an individual rater was 74. The average was 19.

It was also evident that the numbers of arbitrations which had occurred varied greatly from table to table. The lowest number was 36; the highest was 136. The average was 75.9.

Because of the great amount of variability in the data, an important question was that of the degree to which particular raters had influenced the rates for their tables. Since it was not known how many papers each rater or table had read, it was not possible to compute this precisely. However, it was possible to determine the percentage of a table's arbitrations which involved each rater. For certain raters this was found to be quite high.

For 12 of the 24 tables, one individual was involved in at least 40 percent of the arbitrations of the table. For two of these tables two individuals were involved in the great majority of arbitrations. In some cases these two raters had disagreed with each other, in others they had disa-

greed with other table members. At one of these tables, the two raters were involved in 75 out of 86 (87.2 percent) of the arbitrations. At the other table, the two raters were involved in 93 out of 99 arbitrations, or 93.9 percent. For the 12 tables in which no single individual stood out, some raters had somewhat higher numbers of arbitrated papers than did others, but the general pattern was one of having arbitrations divided approximately equally among the table members.

It is possible the magnitude of these results was influenced somewhat by the fact that some raters and tables read more papers than did others. However, the general result is clear. For whatever reason, certain raters were involved in a high proportion of their table's arbitrations; others were not.

Generally those who were involved in large numbers of arbitrations showed this pattern throughout the rating process. Of the 13 who had been involved in at least 40 percent of their table's arbitrations, 11 were involved in at least 30 percent of these throughout each one-third of their table's arbitrations.

Relationships Between Arbitration Results for Each Grade Level

As a means of summarizing a portion of the results for the 189 raters, correlations were performed involving three types of variables. These were the numbers of arbitrations for each rater, the percentage of time the rater gave the higher of the two scores which were arbitrated and the mean of the ratings given by the rater for those papers he or she had read which were arbitrated. These results were computed both overall and for papers read for each of the two grade levels. They are contained in Table 6.

It can be seen first in this table that for each of the three types of measures there were significant correlations between the grade 6 and grade 9 results for the same rater. The numbers of grade 6 and grade 9 arbitrations for each rater correlated .49; the percentage of time a rater gave the higher score in an arbitration correlated .26 for the two levels; and the mean of the grade 6 and

Table 6
Correlations Among Numbers of Arbitrations, Percent of Arbitrations for Which the Rater Gave the Higher Score and Means of Ratings Given

		G6arbs	G9arbs	Totarbs	G6%hi	G9%hi	Tot%hi	G6mean	G9mean	Totmean
Number of Arbitrations	Grade 6 (G6arbs)	1.00								
	Grade 9 (G9arbs)	.49**	1.00							
	Total (Totarbs)	.87**	.86**	1.00						
Percent of Arbitrations for Which Rater Gave Higher Score	Grade 6 (G6%hi)	-.07	-.10	-.10	1.00					
	Grade 9 (G9%hi)	-.02	-.06	-.05	.26**	1.00				
	Total (Tot%hi)	-.05	-.11	-.09	.77**	.76**	1.00			
Mean of Ratings Given	Grade 6 (G6mean)	-.03	-.06	-.05	.78**	.27**	.66**	1.00		
	Grade 9 (G9mean)	-.04	-.04	-.05	.21**	.85**	.63**	.29**	1.00	
	Total (Totmean)	-.13	.00	-.08	.62**	.66**	.82**	.79**	.75**	1.00

* Significant LE .05 ** Significant LE .01

grade 9 ratings which were given correlated .29. These relationships were all positive but certainly not strong. Raters apparently functioned somewhat similarly in scoring the two grade levels, but not in exactly the same way.

It is also evident in the table that, as would be expected, those raters who gave the higher score of those being arbitrated were also those who had the highest mean ratings. These correlations were .78 and .85 for grades 6 and 9, respectively.

Finally, it can be seen that the number of arbitrations for a rater was not related to giving high scores. Raters with high rating tendencies, raters with low rating tendencies and raters who appeared to have no major tendency toward either high or low ratings all were candidates for large numbers of arbitrations.

Relationships Between Rater Characteristics and Arbitrations

Table 7 summarizes the final set of analyses of the study. This table is concerned with the extent to which characterstics of the raters were related to both the frequency of their arbitrations and the magnitude of their ratings.

To prepare Table 7, information about each rater was obtained, including whether he or she was a teacher, his or her gender, whether the rater was employed in an elementary school, an intermediate school or a high school and whether the rater had had previous experience as a scorer in a writing assessment. In addition, for each rater's district, writing means were computed for the grade 6 and grade 9 students who were involved in the writing assesssment. This was to determine the degree to which the achievement level of the students the rater interacted with daily might have influenced his or her ratings.

Table 7
Correlations Between Rater Characteristics
and Both Arbitration Results and Mean Ratings

Characteristic (1)	Number of Arbitrations			Percent of Arbitrations for Which Rater Gave Higher Score			Mean of Ratings Given		
	Grade 6	Grade 9	Total	Grade 6	Grade 9	Total	Grade 6	Grade 9	Total
Position	-.05	.01	-.02	.07	-.06	.05	-.01	-.05	-.01
Sex	-.09	-.03	-.07	.00	-.02	.00	-.05	-.05	-.03
Elementary Experience	-.07	-.14	-.12	-.03	.18*	.08	-.05	.10	.02
Intermediate Experience	-.02	.10	.05	-.01	-.14	-.10	.10	-.04	.06
High School Experience	.08	.01	.05	.02	-.07	-.02	.00	-.06	-.05
Past Rating Experience	-.03	-.03	-.04	.03	.03	.02	-.05	.08	.01
District Grade 6 Writing Mean	-.06	.10	.02	.15*	.05	.14	.15*	.10	.19*
District Grade 9 Writing Mean	-.05	.17*	.07	.10	.01	.09	.11	.04	.15*

* Significant LE .05 ** Significant LE .01

(1) The number of scores entering into each correlation was 189 for the majority of the variables included. However, for the Elementary, Intermediate and High School Experience variables, data were available for only 169 persons. In addition, because not all raters were from districts which participated in the writing assessment, the numbers of scores available for the District Writing Mean variables were 144 and 140 for grades 6 and 9, respectively. The characteristics listed above were coded numerically. Codings for each variable were as follows: Position--1 = Teacher; 2 = Non-teacher working within a school; 3 = Educator but not working in a school. Sex--1 = Male; 2 = Female. Elementary Experience--1 = Working within a building housing grades k-6; 0 = Not working in an elementary building. Intermediate Experience--1 = Working in a building housing grades 7-8; 0 = Not working in an intermediate building. High School Experience--1 = Working in a building housing grades 9-12; 0 = Not working in a high school building. Past Rating Experience--1 = Served as a rater scoring a writing assessment before; 0 = Never served as a rater in a writing assessment. District Grade 6 Writing Mean--Writing mean of grade 6 students from rater's district who were involved in the assessment. District Grade 9 Writing Mean--Writing mean of grade 9 students from rater's district who were involved in the assessment.

As is evident in Table 7, no strong relationships were found between rater characteristics and either arbitration frequencies or the magnitude of raters' scorings. There were, however, a few significant correlations existent.

First of all, working in an elementary school correlated .18 with the percentage of time the rater gave the higher of the two scores in an arbitration involving a grade 9 paper. It is possible this was a function of seeing the grade 9 papers as more competent since these raters, who were primarily elementary teachers, were daily exposed to a much lower level of writing quality.

The other significant correlations in the table all involved the district writing mean variables. Raters whose grade 6 students wrote higher-rated essays tended also to give higher ratings to grade 6 papers. And, raters whose grade 9 students wrote higher-rated essays tended to be involved in more grade 9 arbitrations but also tended to give higher scores in general. These results are opposite to those which might have been expected if raters were being influenced in their scoring by the past experiences they had had in reading their own students' essays. It would have been expected that those with high achieving students might have scored the essays of the assessment lower in comparison with those they were used to seeing in their classrooms.

DISCUSSION AND CONCLUSIONS

The study of arbitrations in Pennsylvania's writing assessment provided a much more in-depth look at the process than had been available previously. In general, support was found for the training which had taken place. The overall arbitration rate of five percent indicates that the vast majority of the papers which were processed were scored with little disagreement. Most raters had obviously internalized the scoring scale very well.

One finding of the study was that arbitration rates differed by prompt, by mode and across the two grade levels. It is not clear why this was the case, but the data collected provided the basis for some tentative conclusions.

First of all, as might be expected, extreme scores given by raters tended most often to be arbitrated. There were differences across grade levels in this respect, however. At the grade 9 level both high and low extreme scores had a similar likelihood of being arbitrated. But at the grade 6 level there was much more likelihood of high scores being arbitrated than low scores.

The results seem to indicate that many arbitrations may be due to raters "testing the limits" of the scoring system. There do appear to be limits upon ratings which can reasonably be given with little disagreement. Grade 6 students may have good, creative ideas but, because of their inexperience in formulating these into a coherent essay, a rating at the top of the scale is one with a high probability of being sent to arbitration. At this level, not only were a high percentage of "6" scores arbitrated, but also the arbitration rate was highest for those prompts for which the highest mean ratings were given.

At the grade 9 level there was no clear pattern evident. Both prompts for which ratings were high and prompts for which ratings were low were among those with the highest arbitration rates. This suggests that, after a certain level of writer maturity has been reached, the deciding factor in terms of arbitrations is more likely to be the prompt itself than it is some general factor.

Although the study did not in any way attempt to analyze differences in the prompts themselves, the data do suggest one hypothesis. For two of the prompts the arbitration rates at both grade levels were among the highest. For one prompt the arbitration rates for both grade levels were among the lowest. By comparing these high and low arbitrated prompts, a conclusion is suggested.

The two prompts with high arbitration rates were number 3, "Time Machine," in which writers were to speculate about the future, and number 8, "Advertising," in which they were to persuade an adult to purchase for another student an item the writer enjoyed. The prompt with the low rate was number 6, "Holiday Explained," for which students were to explain some holiday of their choice.

These prompts with high and low arbitration levels seem to differ in the degree to which it is difficult for a student to write a focused essay. The holiday prompt is very straight-forward. All students have experienced holidays. They simply must choose one and focus their essay upon explaining it.

For both of the prompts with high arbitration levels, there is a degree of probability that the writer might not really have a great deal to say. He or she may not have much idea about the future. He or she may have no particular possession to recommend as strongly as is required. The focus of such an essay must suffer, even if a student is an adept writer. This type of essay would seem to be especially problematical when a holistic scale is used--students may deserve a "6" for some aspects of their writing but a "4" for others, thus providing a situation leading to rater disagreements.

The above hypothesis was stated in an attempt to make some sense out of the differences in arbitration rates computed in the study. It is merely speculation at this point. Research is needed which addresses the question of why the use of certain prompts leads to higher arbitration rates. A better knowledge of this would make it possible to avoid topics which lead to these higher rates.

It was also found in the study that approximately one-half of the raters showed some tendency toward giving either the higher or the lower of the two scores that were arbitrated. These data support the contention of Hake (1986) and of Webb, et al (1990) that raters tend to differ in terms of the leniency or stringency of their own internal scales. They also support the procedures which have routinely been instituted in many assessment programs calling for a balancing of raters such that particular schools or individuals are not penalized by having their papers scored by an especially severe rater.

In addition to examining raters' general tendencies to give either high or low scores, other factors, such as the grade level and writing achievement level of the raters' students were investigated. If the training process is not functioning properly one would expect raters to be operating on the basis of factors other than the scale they have been taught. That only weak relationships were

found testifies to the success of the training in preventing these factors from becoming major influences upon raters' scores.

Finally, at one-half of the tables of raters there were individuals who were involved in a high percentage of the table's arbitrations. This was true throughout the scoring session even though recalibrations were carried out a number of times. From the data available it was not possible to determine the degree to which the high numbers of arbitrations for these raters were primarily due to their inability to conform to the holistic scale and the degree to which it was due to other factors, such as their reading papers at a faster rate than others. In any case, there is little doubt certain raters were detriments to the efficiency of the process. These results argue for asking raters who cannot internalize the scale to depart, especially in "high-stakes" situations.

Thus, the study focused upon an important aspect of measurement today, that of the error which takes place in a performance assessment. As the movement in the country toward more and more use of performance assessment continues, the need to increase the reliability of such measures becomes increasingly important. If we are to objectify the scoring of not only writing assessments but also other types of performances, we must understand the factors which enter into making such measures either reliable or fraught with error. This study was one attempt to understand this type of error. Many more similar investigations are necessary.

REFERENCES

Daly, J.A. & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19*, 309-316.

Diederich, P.B. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Freedman, S.W. (1981). Influences of evaluation of expository essays: Beyond the text. *Research in the Teaching of English, 15*, 245-255.

Hake, R. (1986). How do we judge what they write? In K.L. Greenberg, H.S. Wiener, & R.A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 153-167). New York: Longman.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

Meredith, V.H. & Williams, P.L. (1984). Issues in direct writing assessment: Problem identification and control. *Educational Measurement: Issues and Practice,3*, 11-15, 35.

Peters, L. & Miller, L. (1990). *Writing assessment handbook.* Harrisburg, PA: Pennsylvania Department of Education.

Quellmalz, E.S. (1990). Toward successful large-scale writing assessment: Where are we now? Where do we go from here? *Educational Measurement: Issues and Practice, 3* 29-32, 35.

Rigsby, L.C. (1987, March). Changes in students' writing and their significance. Paper presented at the meeting of the Conference on College Composition and Communication, Atlanta.

Stock, P.L. & Robinson, J.L. (1987). Talking on testing: Teachers as testers, researchers. *English Education,19*, 93-121.

Webb, L.C., Raymond, M.R. & Houston, W.M. (1990, April). Rater stringency and consistency in performance assessment. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

APPENDIX

PENNSYLVANIA WRITING ASSESSMENT
HOLISTIC SCORING GUIDE

6	5	4	3	2	1
• sharp, distinct focus • substantial, specific, and/or illustrative content; sophisticated ideas that are particularly well developed • obviously controlled and/or subtle organization • writer's voice apparent in tone, sentence structure, and word choice • few mechanical and usage errors	• clear focus • specific and illustrative content • logical and appropriate organization • precision and variety in sentence structure and word choice • some mechanical and usage errors	• adequate focus • sufficient content • appropriate organization • some precision and variety in sentence structure and word choice • mechanical and usage errors not severe enough to interfere significantly with the writer's purpose	• vague focus • content limited to a listing, repetition, or mere sequence of ideas • inconsistent organization • limited sentence variety and word choice • repeated weaknesses in mechanics and usage	• confused focus • superficial content • confused organization • lack of sentence and word choice variety • mechanical and usage errors that seriously interfere with the writer's purpose	• absence of focus • absence of relevant content • absence of organization • no apparent control over sentence structure and word choice • mechanical and usage errors so severe that writer's ideas are difficult if not impossible to understand

NON-SCOREABLE (NS)
• is illegible: i.e., includes so many undecipherable words that no sense can be made of the response or • is incoherent: i.e., words are legible but syntax is so garbled that response makes no sense or • is a blank paper

CHARACTERISTICS OF EFFECTIVE WRITING

FOCUS	CONTENT	ORGANIZATION	STYLE	CONVENTIONS
• demonstrates an awareness of audience and task • establishes and maintains a clear purpose • sustains a single point of view • exhibits clarity of ideas	• information and details are specific to topic • information and details are relevant to focus • ideas are fully developed	• logical order or sequence is maintained • paragraphs deal with one subject • logical transitions are made within sentences and between paragraphs • introduction and conclusion are evident	• precise language • effective word choice • voice, tone, originality of language • variety of sentence structures, types, and lengths	• mechanics: spelling, capitalization, punctuation • usage (e.g., pronoun references, subject-verb agreement) • sentence completeness

BEST COPY AVAILABLE